CSE574 Introduction to Machine Learning

Clustering

Jue Guo

University at Buffalo

March 29, 2024

Outline

Learning Objective

- Define clustering.
- Calculate cluster centroids and inertia.
- List steps in the k-means clustering algorithm.
- Use the elbow and silhouette methods to select the optimal number of clusters.

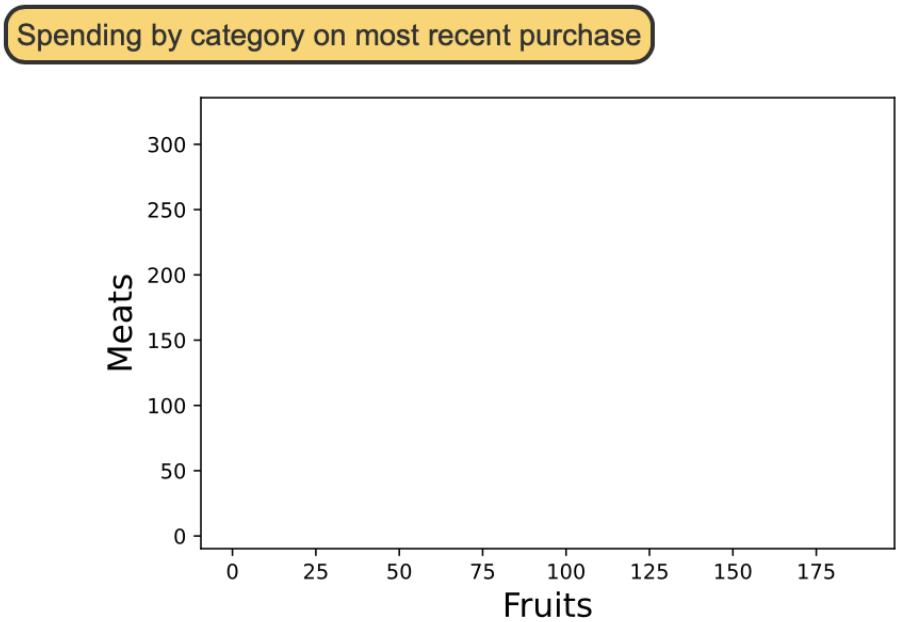
Clustering

Clustering is an unsupervised learning task in which instances are grouped based on similarities in the input features.

- Since clustering is unsupervised, no target output features exist. Instead, clustering results in a new feature containing group assignments.

Clustering algorithms use similarity measures to group instances, such as distance or correlation. Applications of clustering include customer segmentation, recommendation systems, and social network analysis.

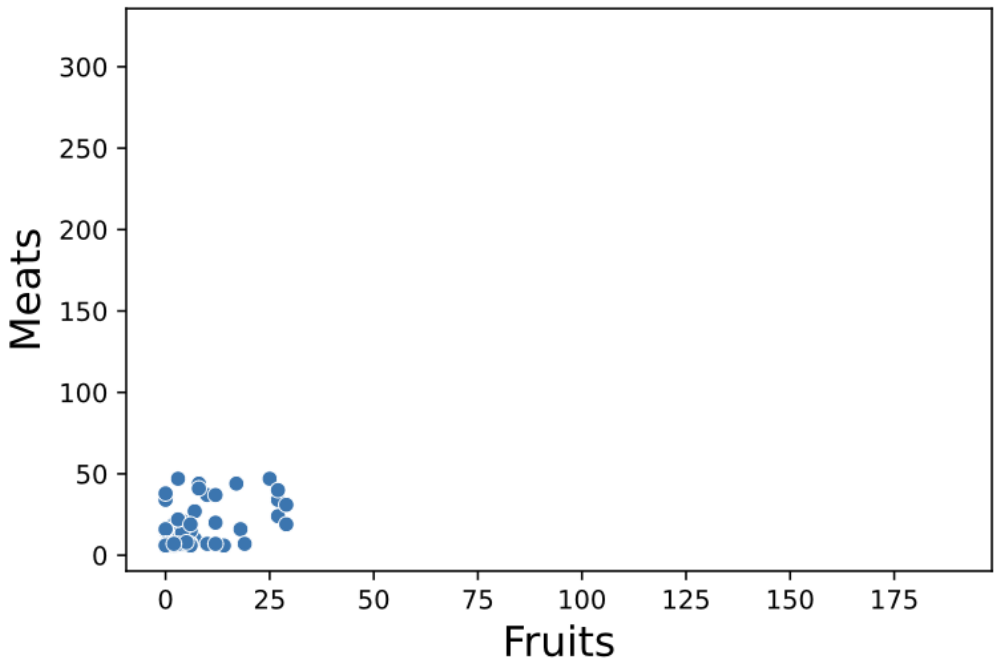
Clustering grocery customers.



Retailers use customer loyalty programs to track spending habits. Ex: A grocery store chain recorded spending by category using customers' most recent purchases.

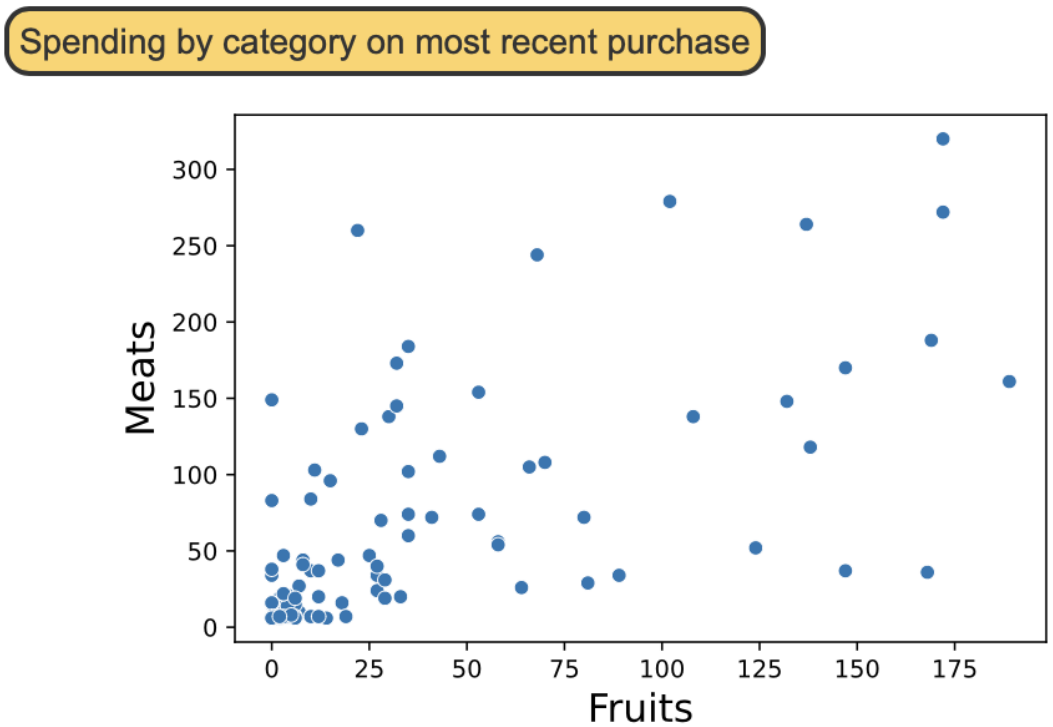
Clustering grocery customers.

Spending by category on most recent purchase



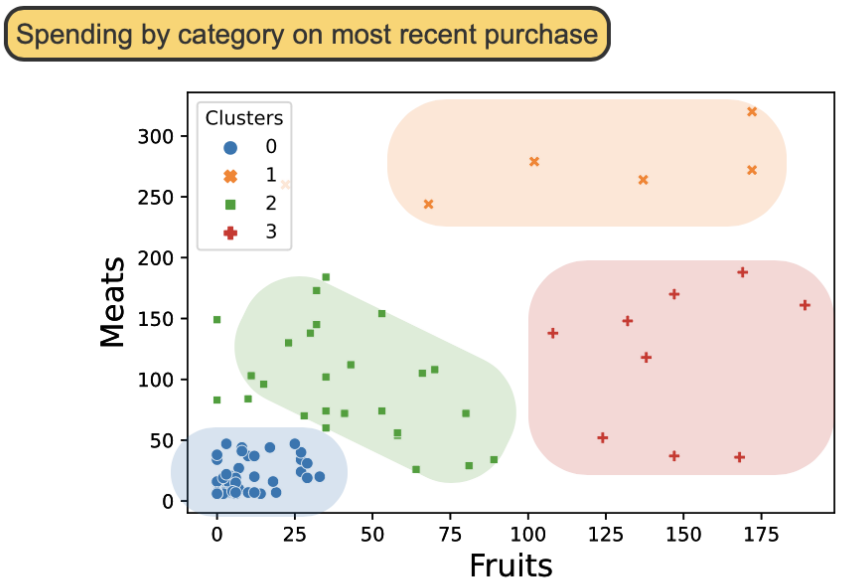
Most customers spent $0-$25 on fruits and $0-$50 on meats.

Clustering grocery customers.



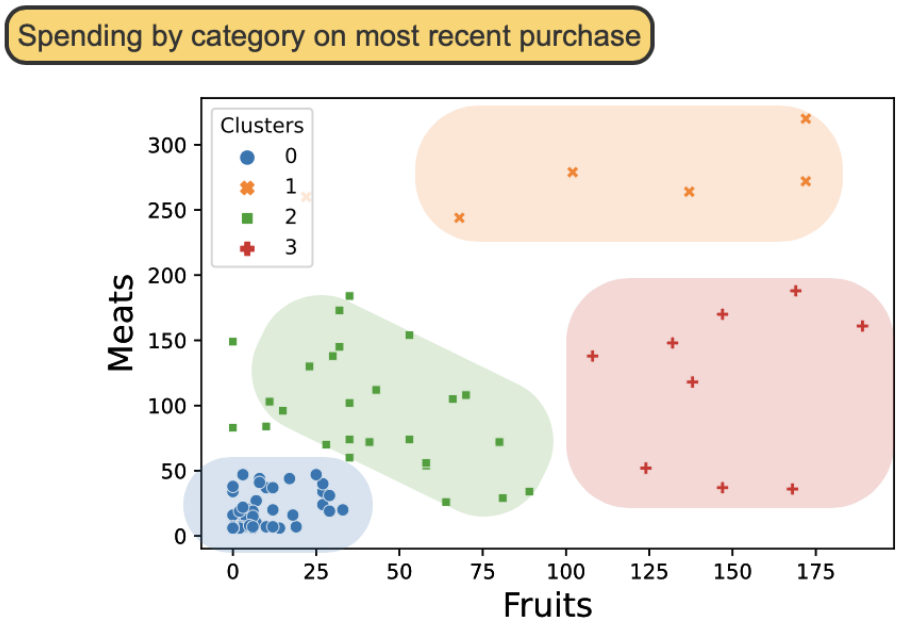
Some customers spent much more on meats, fruits, or both.

Clustering grocery customers.



Clustering groups customers into four categories: low spending, moderate spending, high fruit spending, and high meat spending.

Clustering grocery customers.



Clustering groups customers into four categories: low spending, moderate spending, high fruit spending, and high meat spending.

Cluster centroids

The **centroid** of a cluster is the mean position of the cluster's instances. Centroids are used to summarize the position of the cluster and assign instances to clusters. For a cluster C_i, the centroid's value is:

$$\overline{\mathbf{x}_i} = \frac{\sum_{j \in C_i} \mathbf{x}_j}{n_i}$$

where $\mathbf{x}_j$ is a p-dimensional input vector containing the input feature values for instance j and n_i is the number of instances in cluster k.

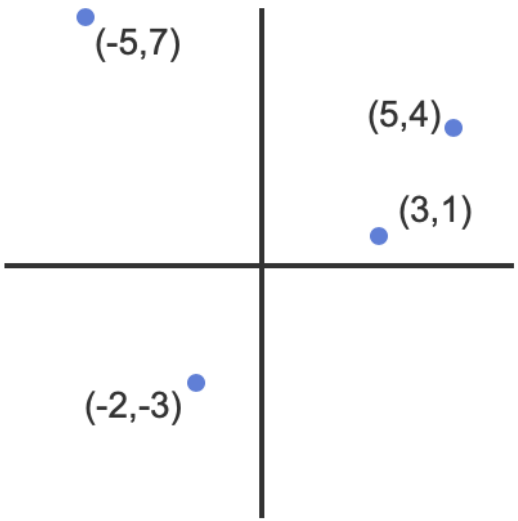
Cluster centroids

The **inertia** of a cluster is the average of each instance's squared distance to the centroid. For cluster i, the inertia is:

$$I_i = \frac{\sum_{j \in C_i} \left| \mathbf{x}_j - \overline{\mathbf{x}_i} \right|^2}{n_i}$$

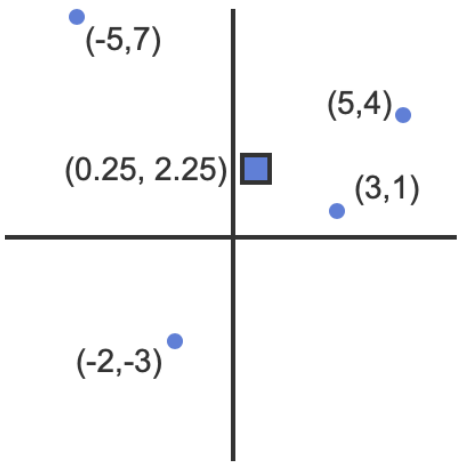
Inertia is also called the cluster's sum of squares. Clusters with low inertia contain more similar instances than clusters with high inertia.

Calculating centroids and inertia.



A cluster has four points: (-5, 7), (5, 4), (3, 1), and (-2, -3).

Calculating centroids and inertia.

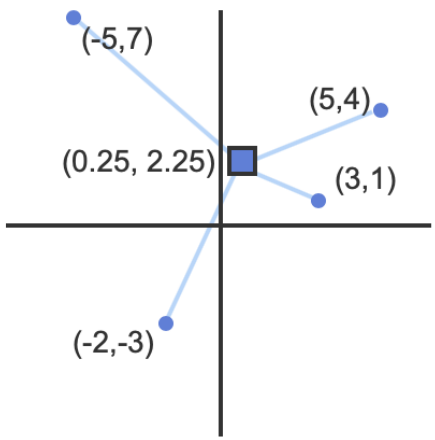


Centroid

$$\overline{\mathbf{x}_k} = \left(\frac{-5+(-2)+3+5}{4}, \frac{7+(-3)+1+4}{4}\right)$$
$$= (0.25, 2.25)$$

The centroid is calculated by averaging each feature.

Calculating centroids and inertia.



Centroid

$$\overline{\mathbf{x}_k} = \left(\frac{-5+(-2)+3+5}{4}, \frac{7+(-3)+1+4}{4}\right)$$
$$= (0.25, 2.25)$$

Inertia

$$I_k = \frac{7.1^2+4.8^2+3.0^2+5.1^2}{4}$$
$$= 27.12$$

The cluster's inertia is calculated by averaging each point's squared distance to the centroid.

k-means clustering

k-means clustering is a clustering algorithm that assigns instances into the cluster with the nearest centroid.

- In k-means clustering, the number of clusters, k, is chosen in advance. k-means clustering is considered stable if the final cluster assignment does not depend on the starting clusters.
- But, stability is not guaranteed, especially when no clear separation between instances exists.

k-means clustering

Randomly select k points for the initial centroids.

1. Assign each instance to the nearest center.
2. Calculate the new cluster centroid.
3. Continue until the cluster centroids do not change or the maximum number of iterations is reached.

Selecting the optimal number of clusters

Two common methods for choosing the optimal value of k for a dataset are the elbow method and silhouette method.

- The **elbow method** graphs the total inertia of the clusters against values of k and chooses the k for which the curve levels off. Since increasing k also increases model complexity, the elbow method finds the k with the best tradeoff between complexity and inertia.

Selecting the optimal number of clusters

The **silhouette method** calculates the silhouette coefficient for each instance, and chooses the k with the highest mean silhouette score.

- For instance j, the silhouette coefficient is

$$S(j) = \frac{\overline{d_{out}(j)} - \overline{d_{in}(j)}}{\max\left(\overline{d_{out}(j)}, \overline{d_{in}(j)}\right)}$$

where $\overline{d_{\text{out}}(j)}$ is the average distance of instance i to the centroid of all other clusters, and $\overline{d_{in}(j)}$ is the average distance of instance i to all other instances in instance i 's cluster. Silhouette coefficients range from -1 to +1 : An instance with a high silhouette coefficient is close to its own centroid, but far from the other centroids.

Silhouette Plots

Silhouette plots graph the silhouette coefficient for each instance, grouped by cluster. Values of k for which some clusters have all below-average silhouette coefficients, or a high proportion of negative silhouette coefficients, should not be used.

In the following figure, cluster 0 is completely below the average silhouette coefficient when k = 2 (left). When k = 4, the clusters are all closer to the average silhouette coefficient, with no instances having a negative silhouette coefficient.

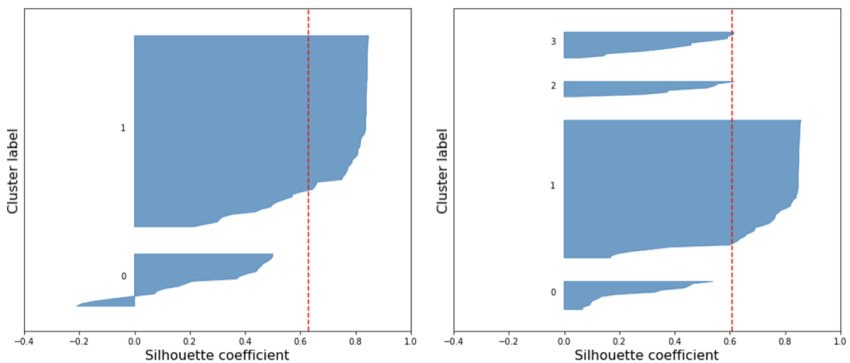
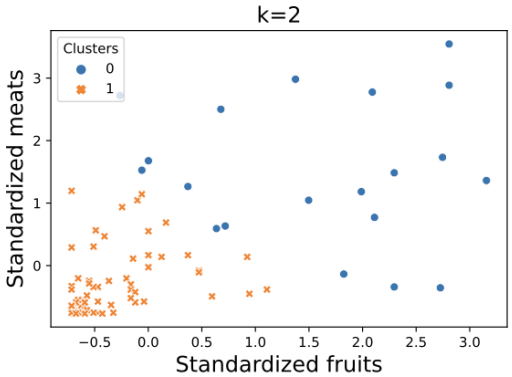


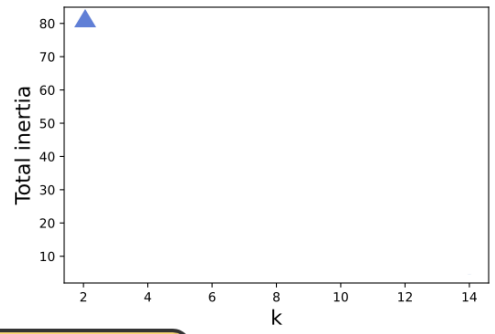
Figure: Silhouette plots for customer spending data.

Selecting the optimal number of clusters.

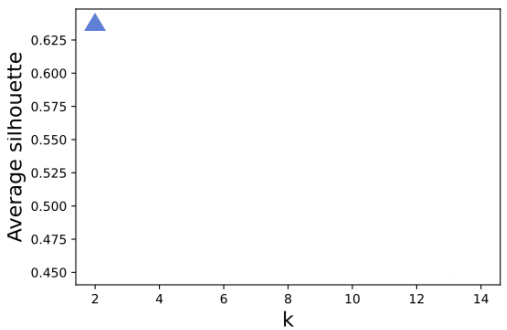
Clusters



Elbow method



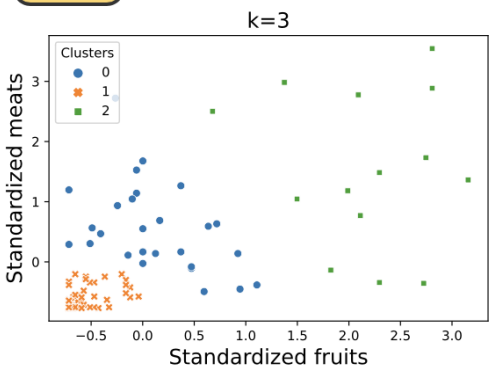
Silhouette method



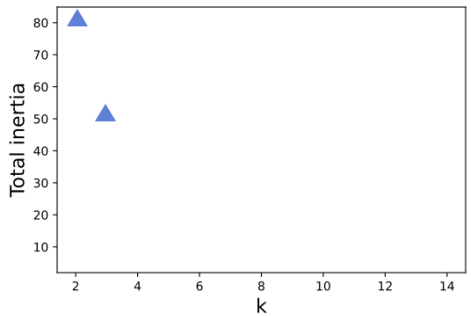
The minimum number of clusters is k = 2. Using two clusters gives inertia = 81.10 and average silhouette = 0.64.

Selecting the optimal number of clusters.

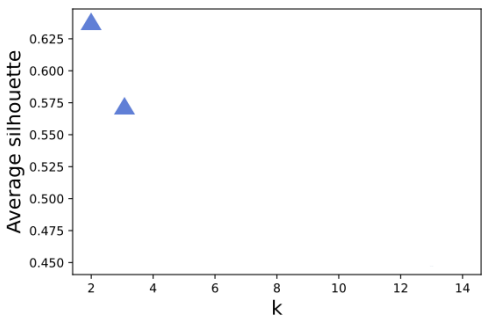
Clusters



Elbow method



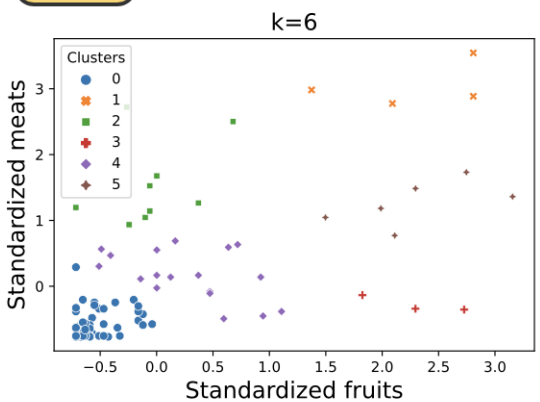
Silhouette method



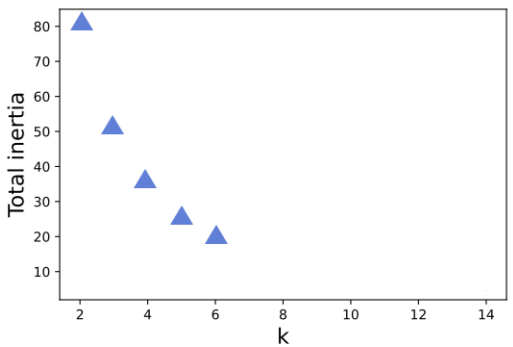
Inertia and average silhouette coefficient are relative measures, so their values are compared against other values of k. Increasing k to 3 decreases both measures.

Selecting the optimal number of clusters.

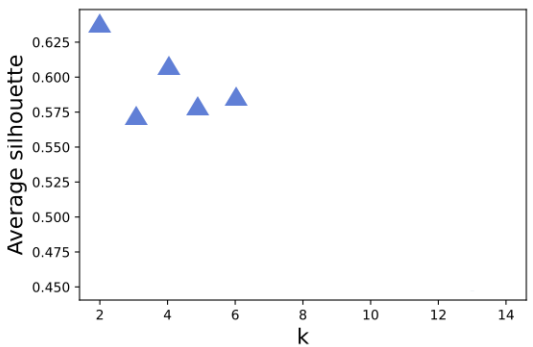
Clusters



Elbow method



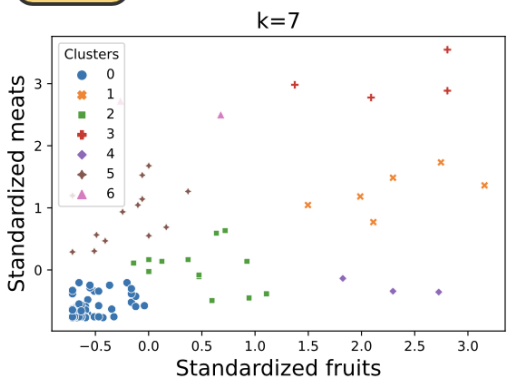
Silhouette method



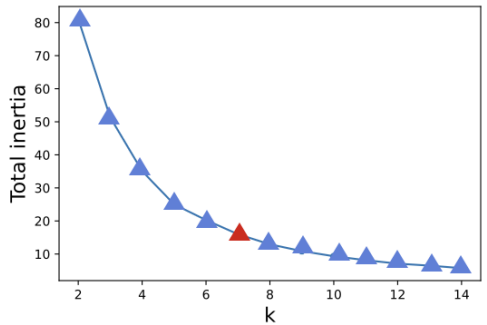
As k increases, inertia and average silhouette tend to decrease.

Selecting the optimal number of clusters.

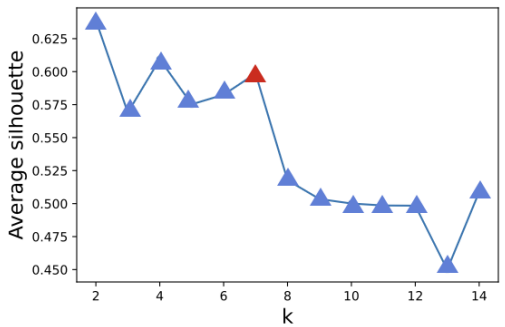
Clusters



Elbow method



Silhouette method



k = 7 is near the leveling off point for inertia and one of the highest average silhouette values, so k = 7 is optimal.

Advantages and Disadvantages of K-means

k-means clustering works well for small and medium sized datasets where all clusters are about the same size. Since instances are assigned to clusters based on distances, all input features should be scaled before clustering.

k-means clustering has several disadvantages:

- *Correlation*: When input features are correlated, k-means clustering fails to identify reasonable clusters. Using principal components instead of the original features can improve interpretability and cluster separation.

Advantages and Disadvantages of K-means

k-means clustering has several disadvantages:

- *Correlation*: When input features are correlated, k-means clustering fails to identify reasonable clusters. Using principal components instead of the original features can improve interpretability and cluster separation.
- *Variance*: k-means clustering assumes all features have the same variance. Standardization or normalization ensures all variances are equal.
- *Cluster sizes*: k-means clustering favors clusters with approximately equal sizes. k-means clustering often groups small natural clusters together.

Advantages and Disadvantages of K-means

- *Shape*: k-means clustering using Euclidean distance prefers clusters with spherical shape. Other distance measures or clustering methods should be used for irregularly shaped clusters.
- *Initial centroids*: The initial centroids directly impact the final clusters.